A quick tutorial on IP Router design

Optics and Routing Seminar
October 10th, 2000

Nick McKeown

nickm@stanford.edu
http://www.stanford.edu/~nickm

Outline

Where IP routers sit in the network

• What IP routers look like

What do IP routers do?

Some details:

– The internals of a "best-effort" router

  • Lookup, buffering and switching

– The internals of a "QoS" router

Can optics help?

Outline (next time)

The way routers are *really* built.

Evolution of their internal workings.

What limits their performance.

- The effect that DWDM is having on switch/router design.

The way the network is built today.

Discussion: The scope for optics

Outline

➡ Where IP routers sit in the network

What IP routers look like

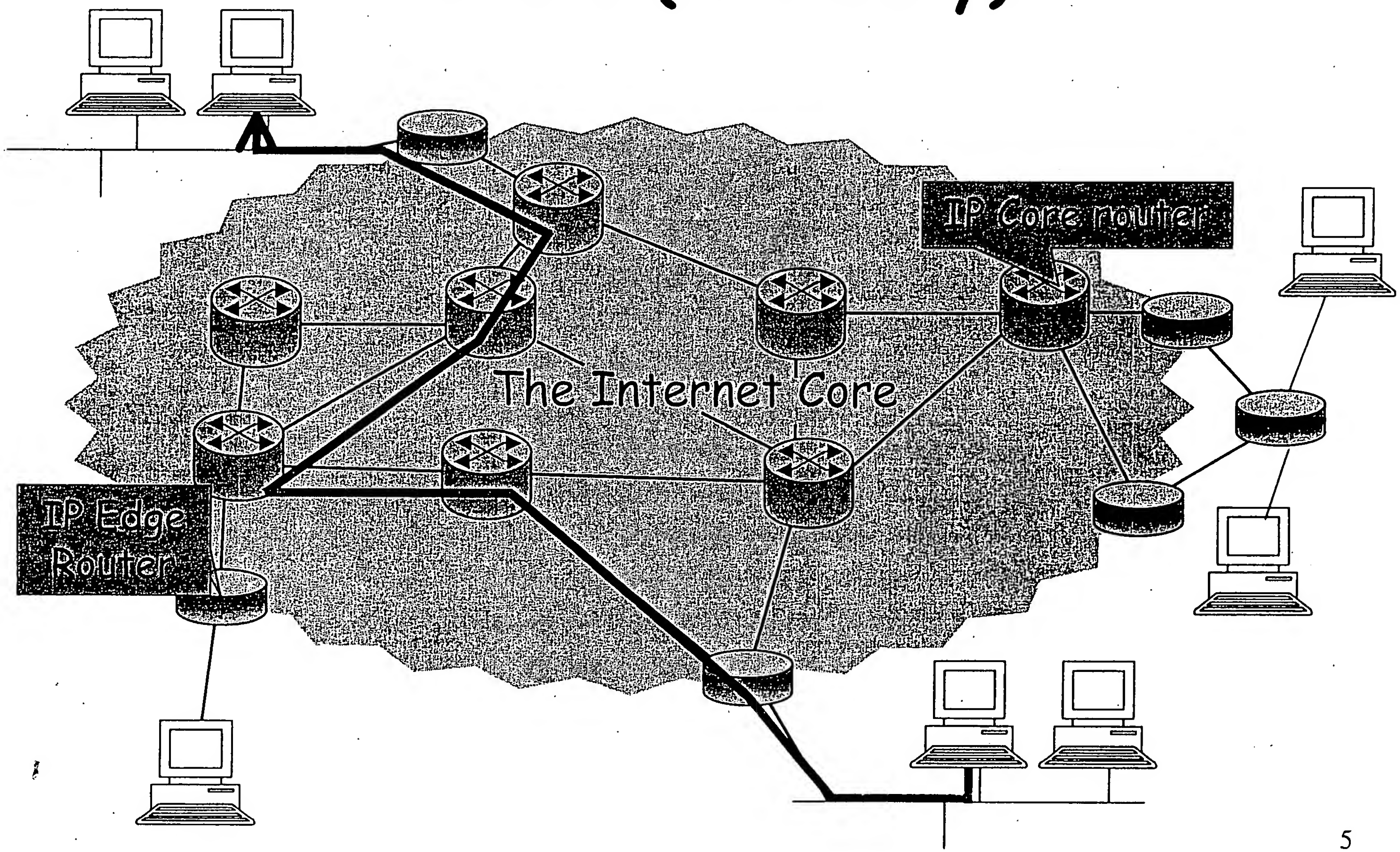
- What do IP routers do?

Some details:

- The internals of a "best-effort" router
    - Lookup, buffering and switching
- The internals of a "QoS" router

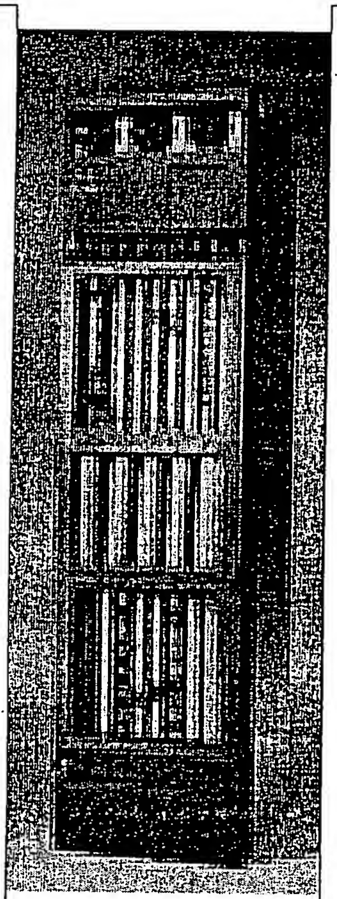
Can optics help?

The Internet is a mesh of routers (in theory)

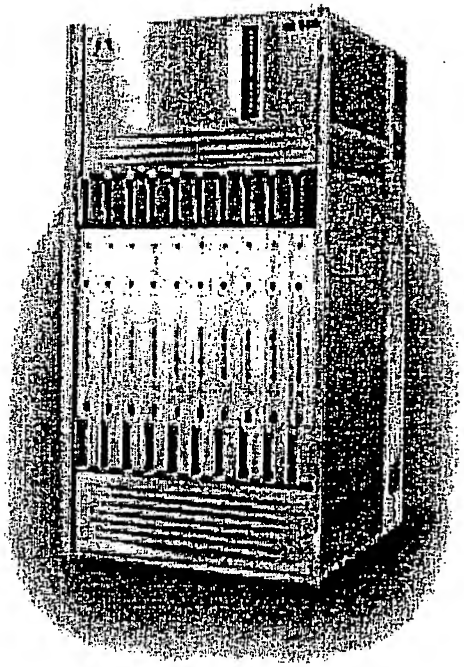


What do they look like?

Access routers
e.g. ISDN, ADSL

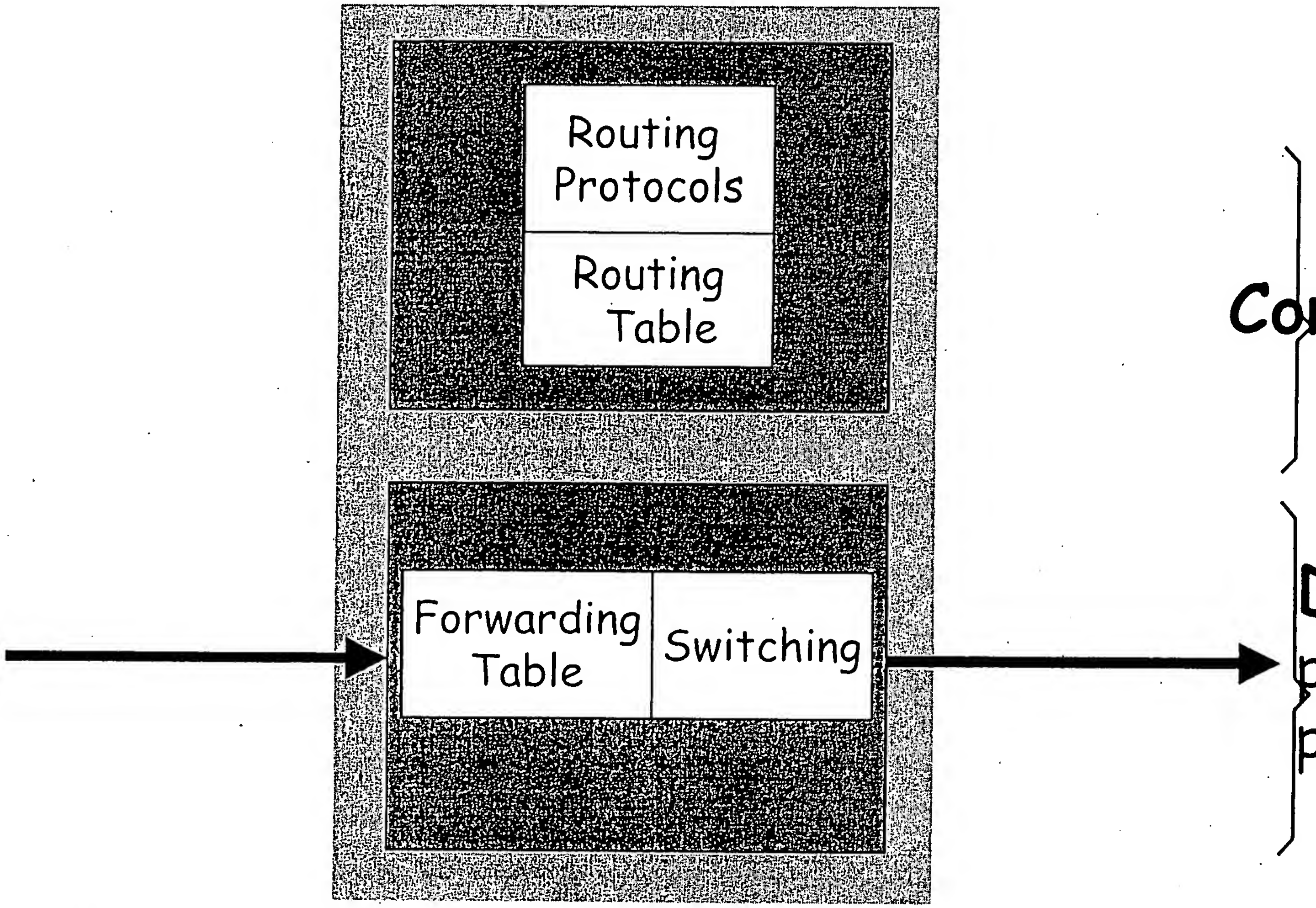


Core router
e.g. OC48c POS



Core ATM switch

Basic Architectural Components of an IP Router



Per-packet processing in an IP Router

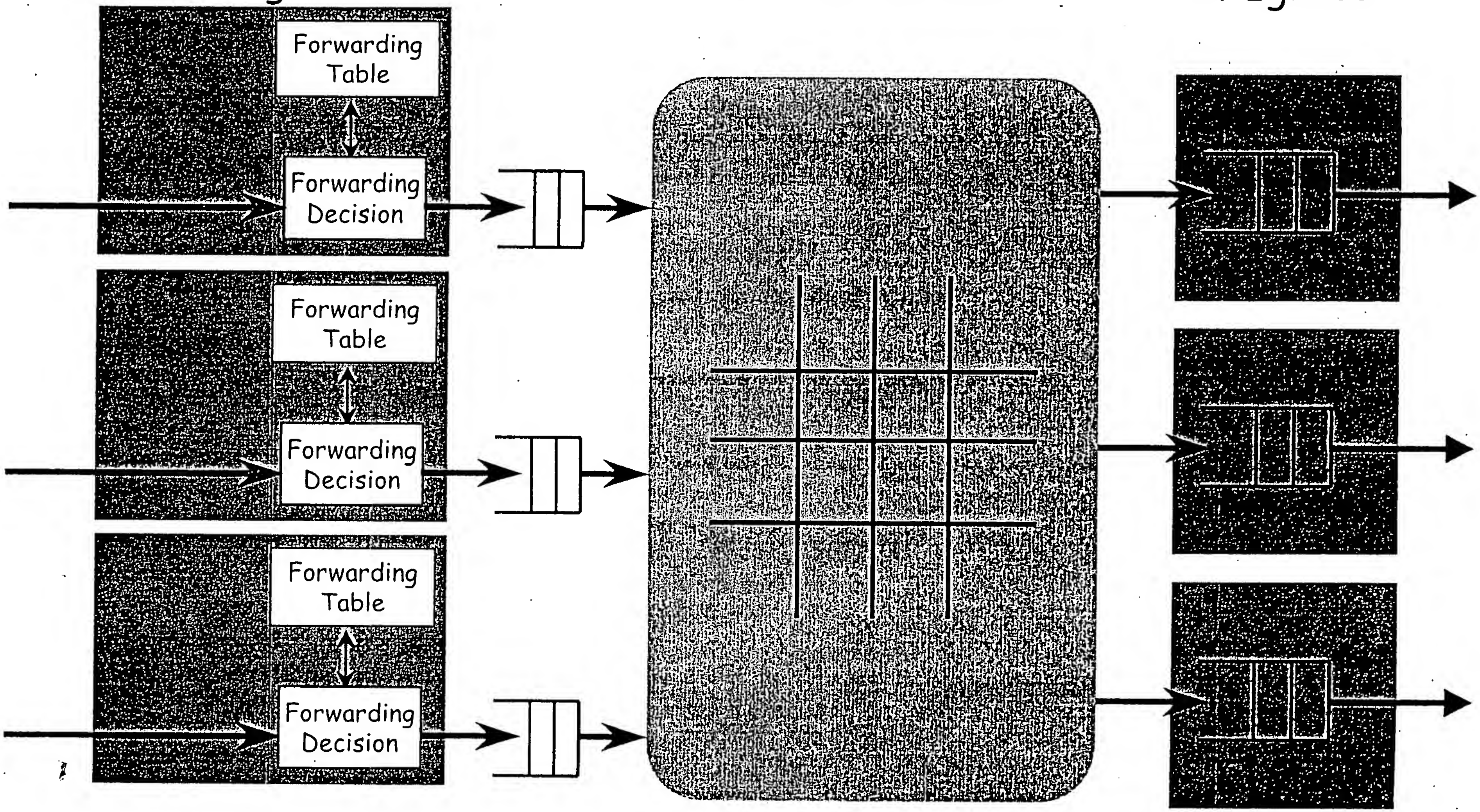
1. Accept packet arriving on an incoming link.
2. Lookup packet destination address in the forwarding table, to identify outgoing port(s).
3. Manipulate packet header: e.g., decrement TTL, update header checksum.
4. Send packet to the outgoing port(s).
5. Buffer packet in the queue.
6. Transmit packet onto outgoing link.

Outline

- Where IP routers sit in the network
- What IP routers look like
- What do IP routers do?
- Some details:
  - The internals of a "best-effort" router
    - Lookup, buffering and switching
  - The internals of a "QoS" router
- Can optics help?

Basic Architectural Components

Datapath: per-packet processing

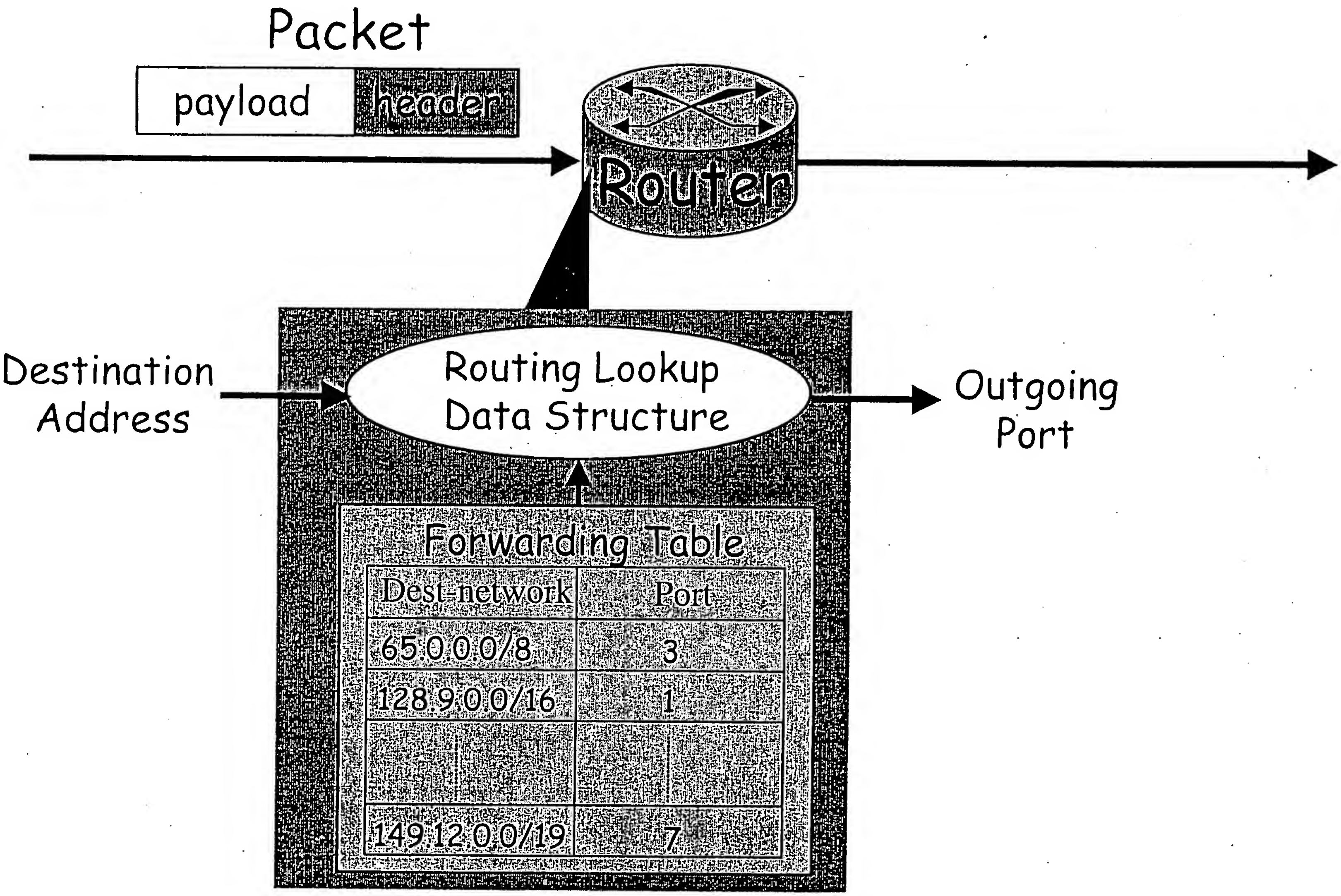


1. Ingress

2. Interconnect

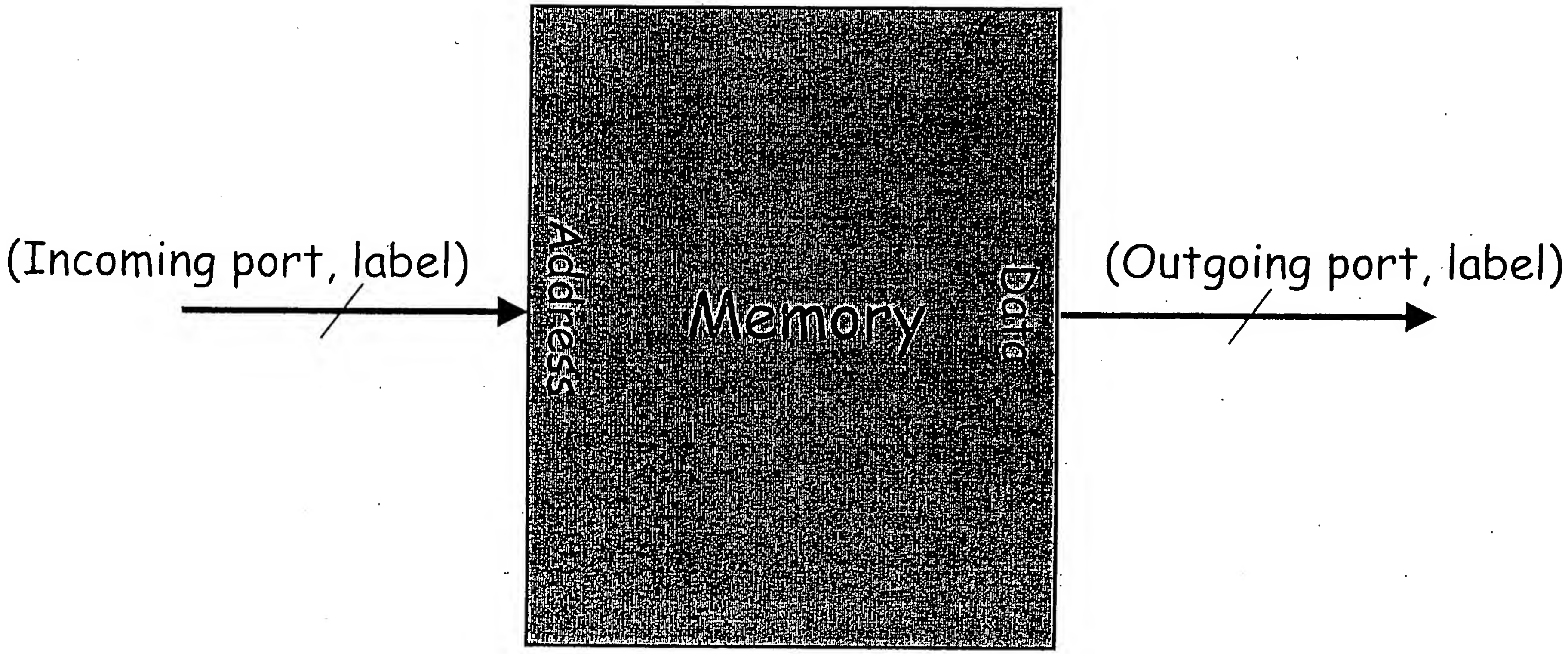
3. Egress

Forwarding Engine



Dest-network	Port
65.0.0.0/8	3
128.9.0.0/16	1
149.12.0.0/19	7

The Search Operation is *not* a Direct Lookup



IP addresses: 32 bits long ⇒ 4G entries

The Search Operation is also not an Exact Match Search

Exact match search: search for a key in a collection of keys of the same length.

Relatively well studied data structures:

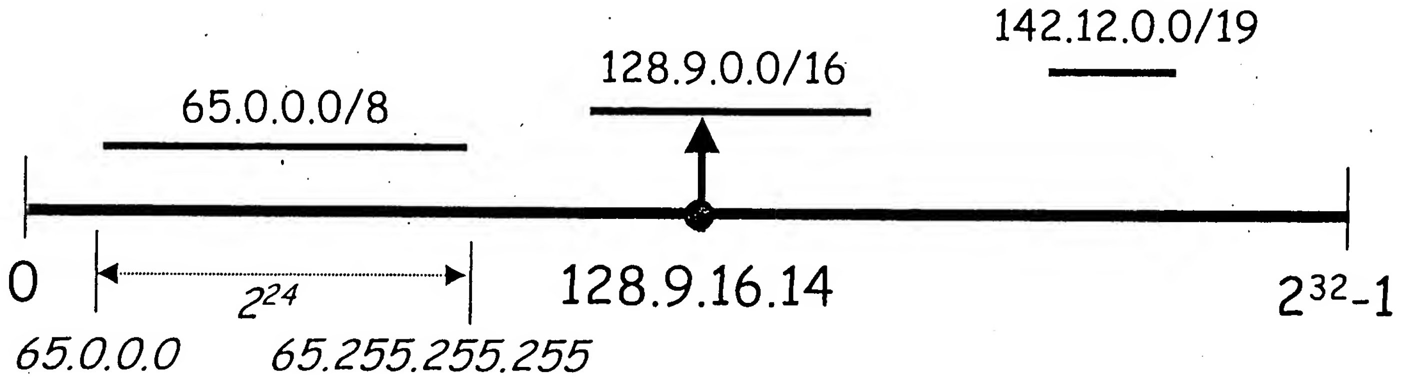
- Hashing
- Balanced binary search trees

Example Forwarding Table

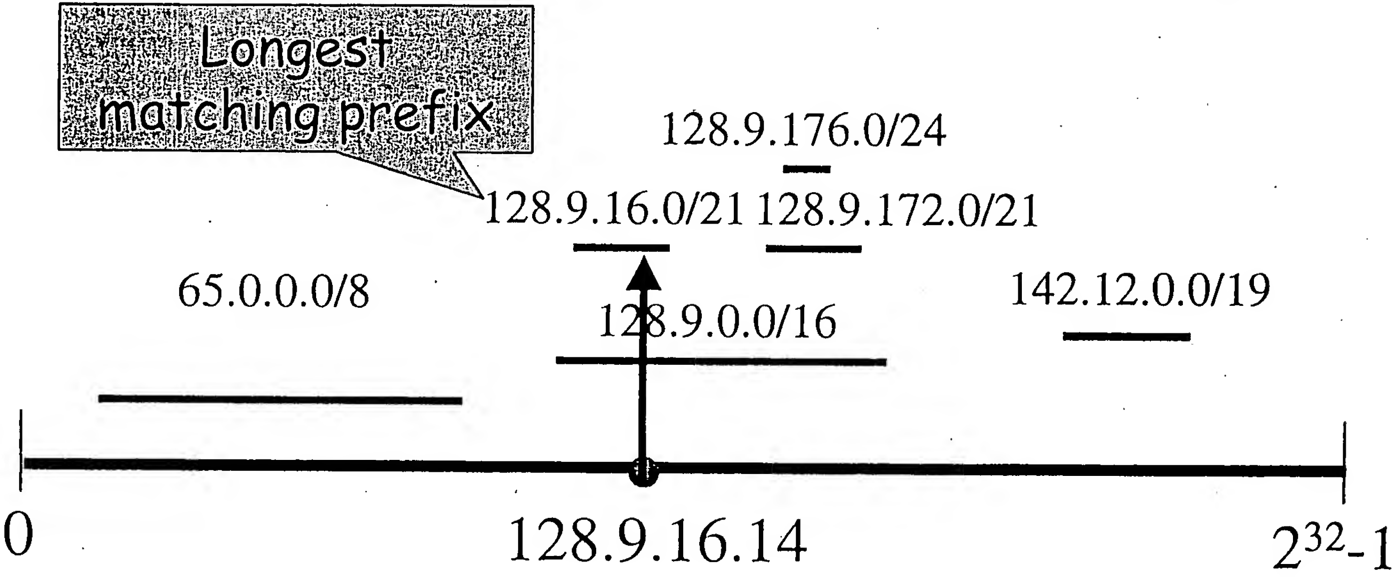
Destination IP Prefix	Outgoing Port
65.0.0.0/8	3
128.9.0.0/16	1
142.12.0.0/19	7

Prefix length

IP prefix: 0-32 bits

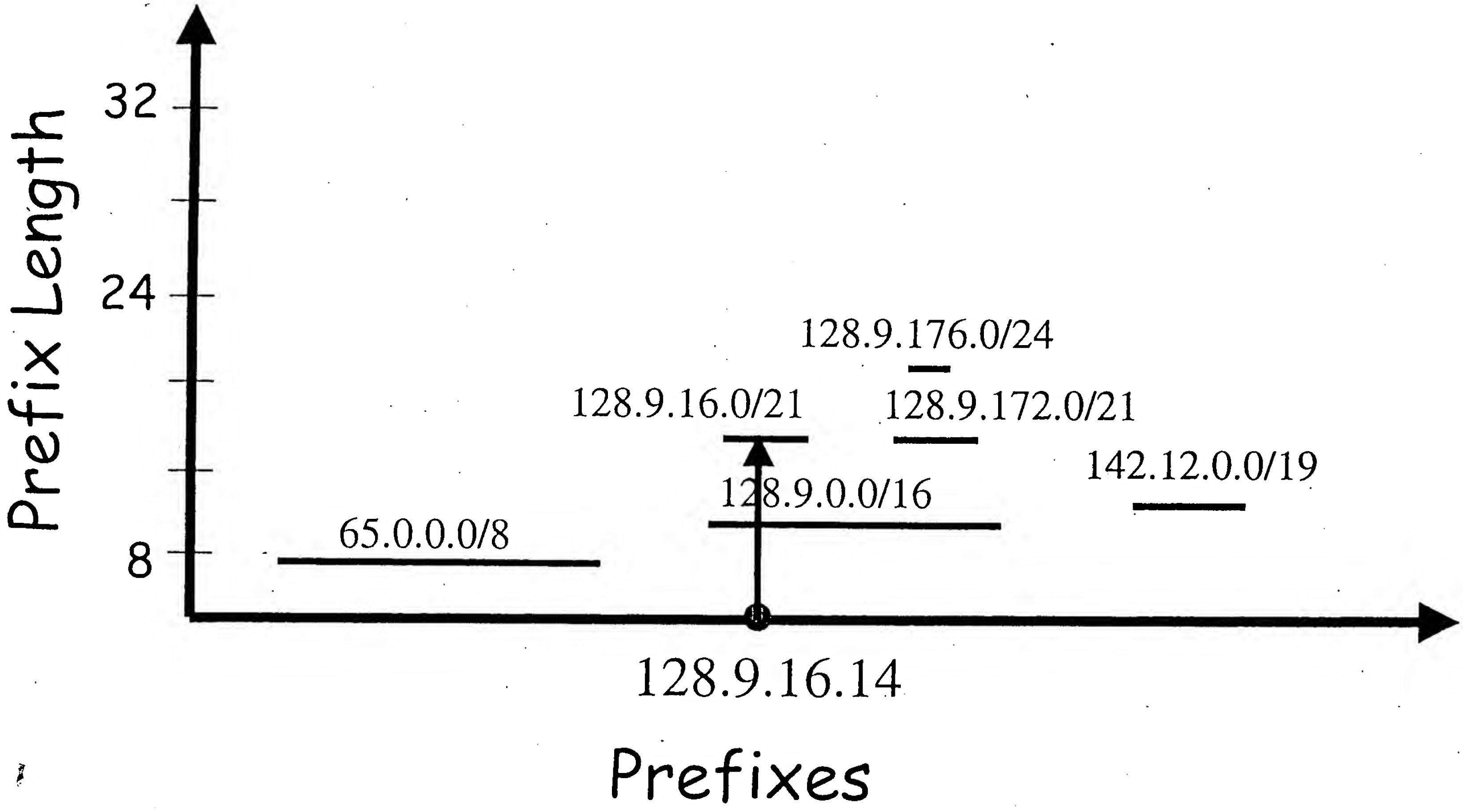


Prefixes can Overlap



Routing lookup: Find the longest matching prefix (aka the most specific route) among all prefixes that match the destination address.

Difficulty of Longest Prefix Match



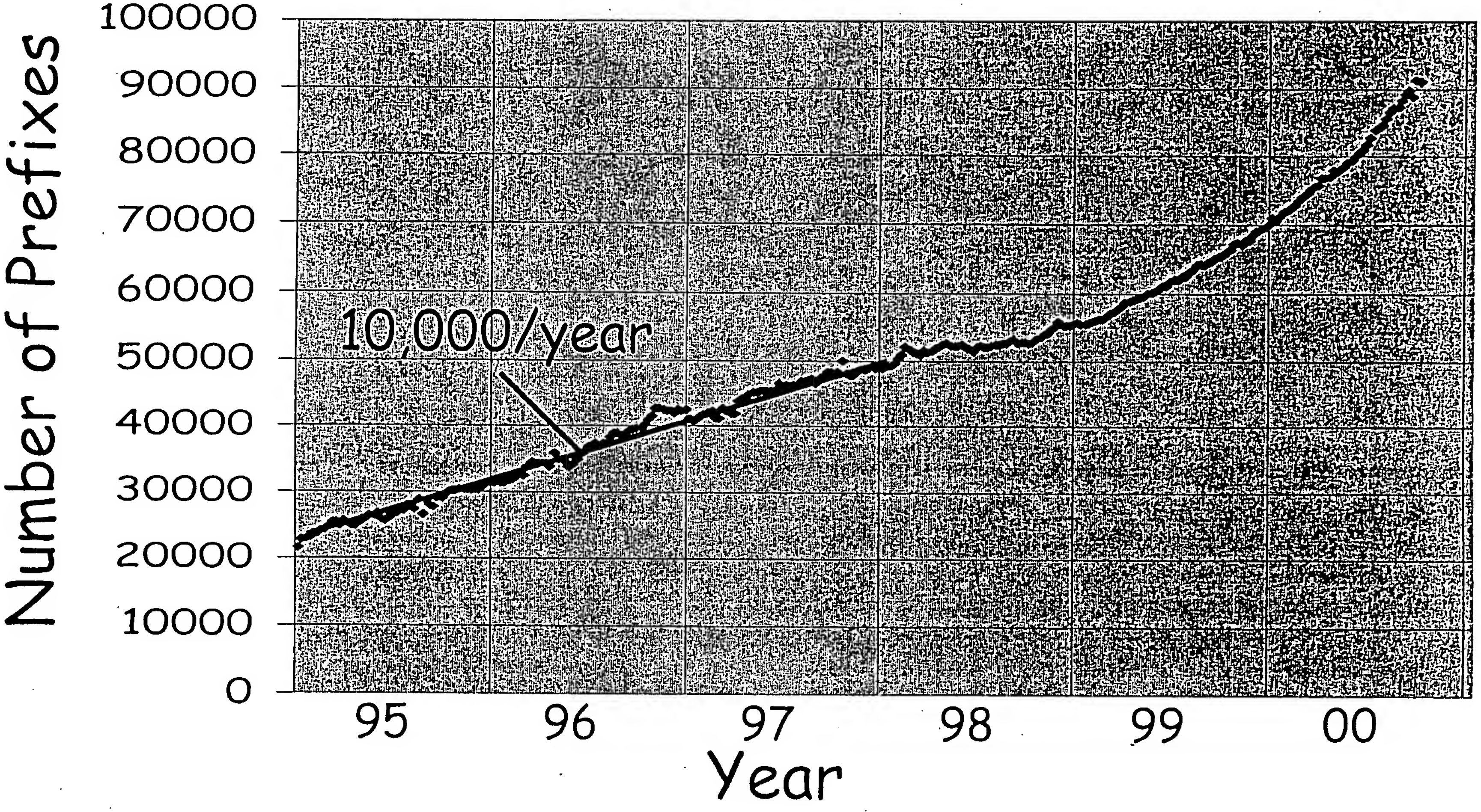
Lookup Rate Required

Year	Line	Line-rate (Gbps)	40B packets (Mpps)
1998-99	OC12c	0.622	1.94
1999-00	OC48c	2.5	7.81
2000-01	OC192c	10.0	31.25
2002-03	OC768c	40.0	125

31.25 Mpps ⇒ 33 ns

DRAM: 50-80 ns, SRAM: 5-10 ns

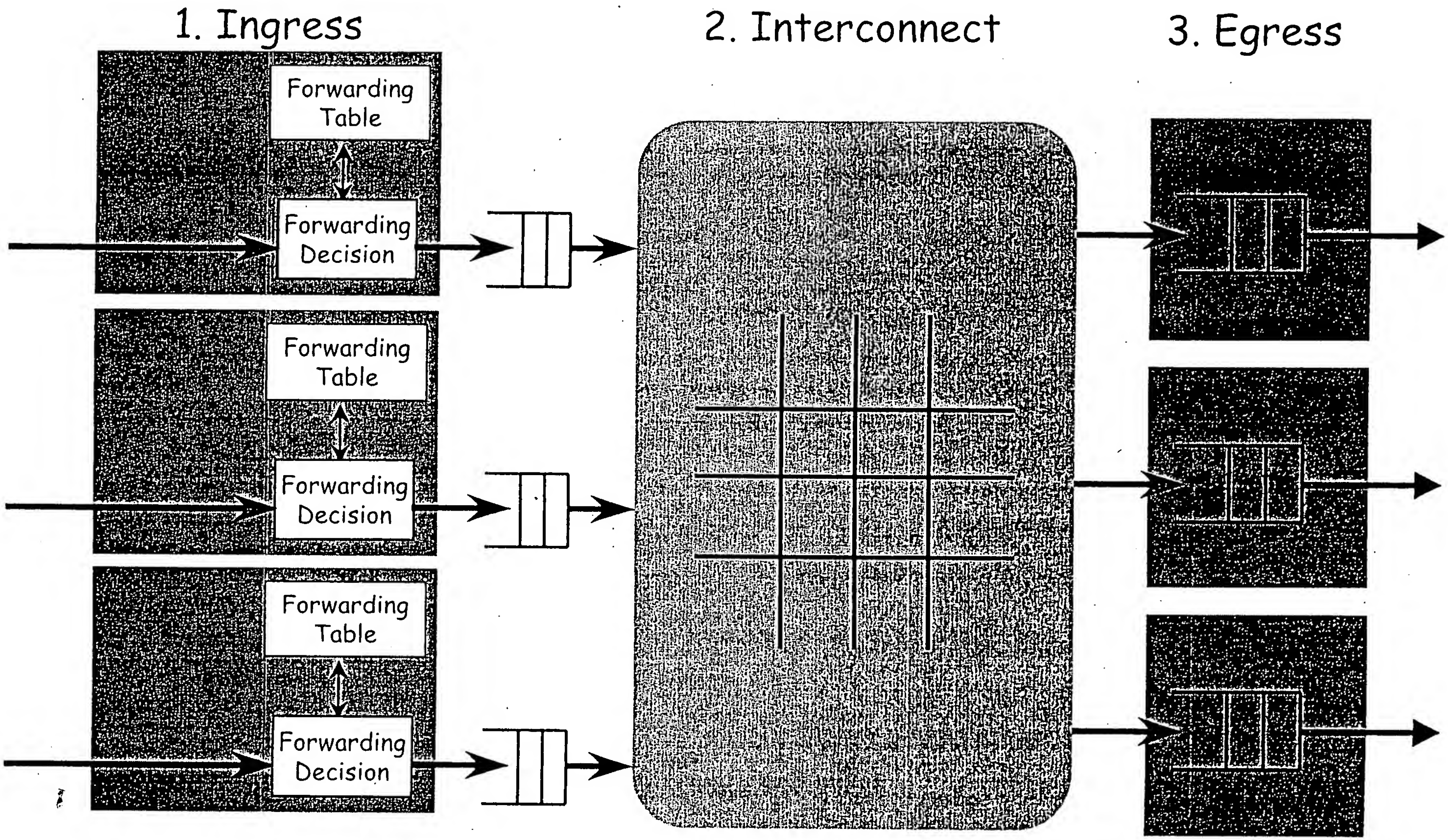
Size of the Forwarding Table



Source: http://www.telstra.net/ops/bgptable.html

Basic Architectural Components

Datapath: per-packet processing



1. Ingress

2. Interconnect

3. Egress

Forwarding Table

Forwarding Decision

Forwarding Table

Forwarding Decision

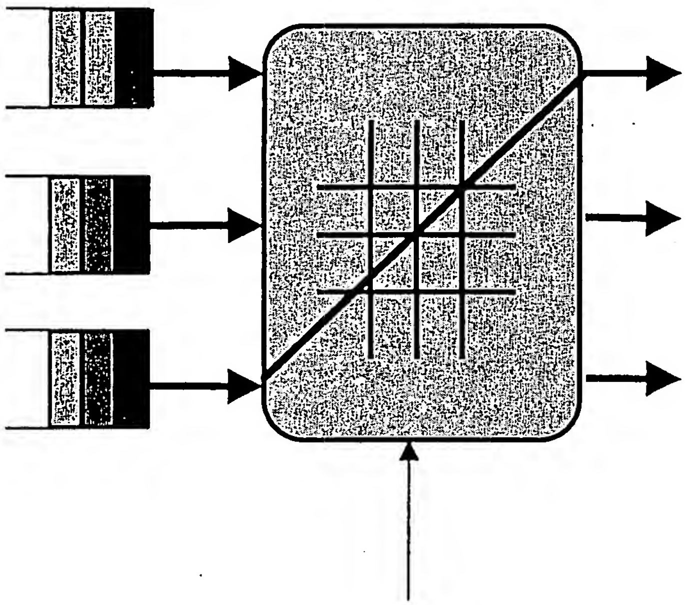
Forwarding Table

Forwarding Decision

Interconnects

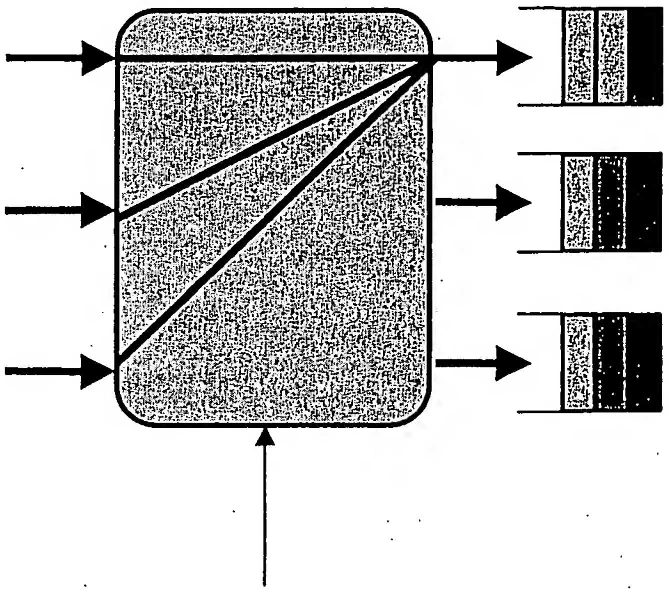
Two basic techniques

Input Queueing



Usually a non-blocking switch fabric (e.g. crossbar)

Output Queueing

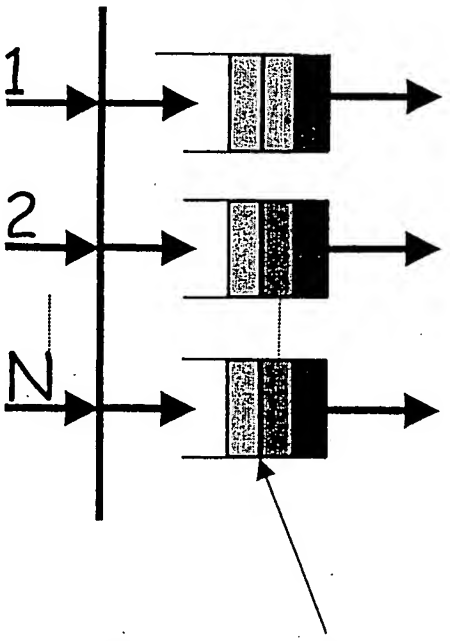


Usually a fast bus

Interconnects

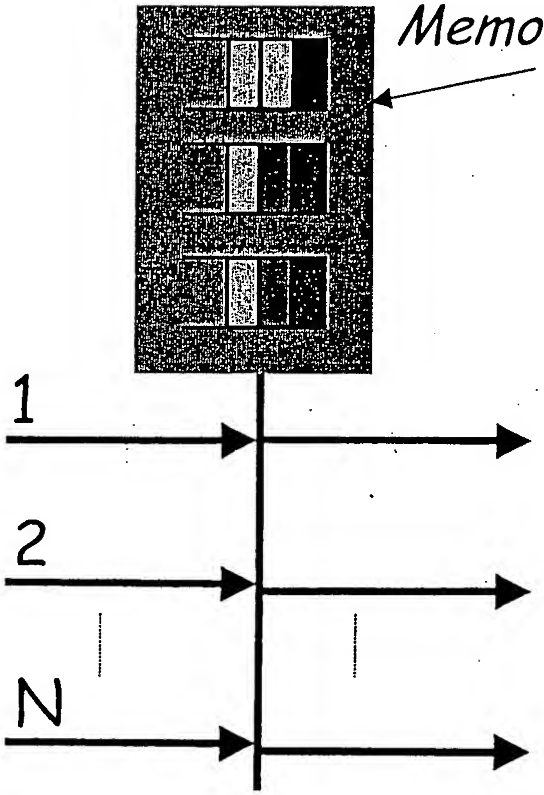
Output Queueing

Individual Output Queues



Memory b/w = (N+1).R

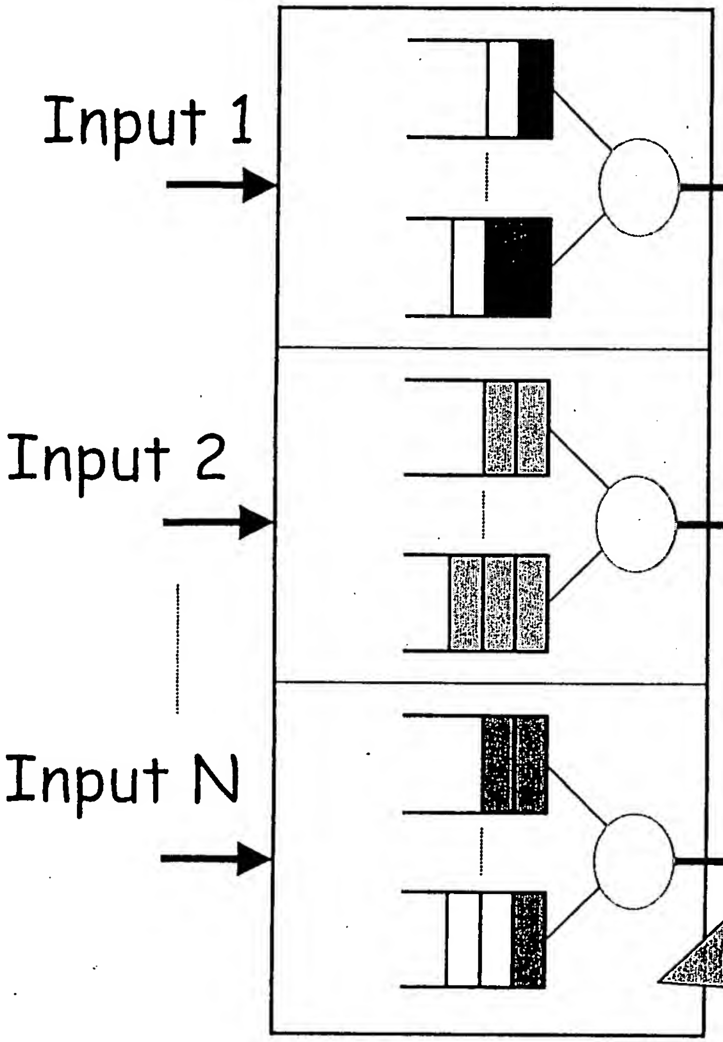
Centralized Shared Memory



Memory b/w = 2N.R

Interconnects

Centralized Shared Memory



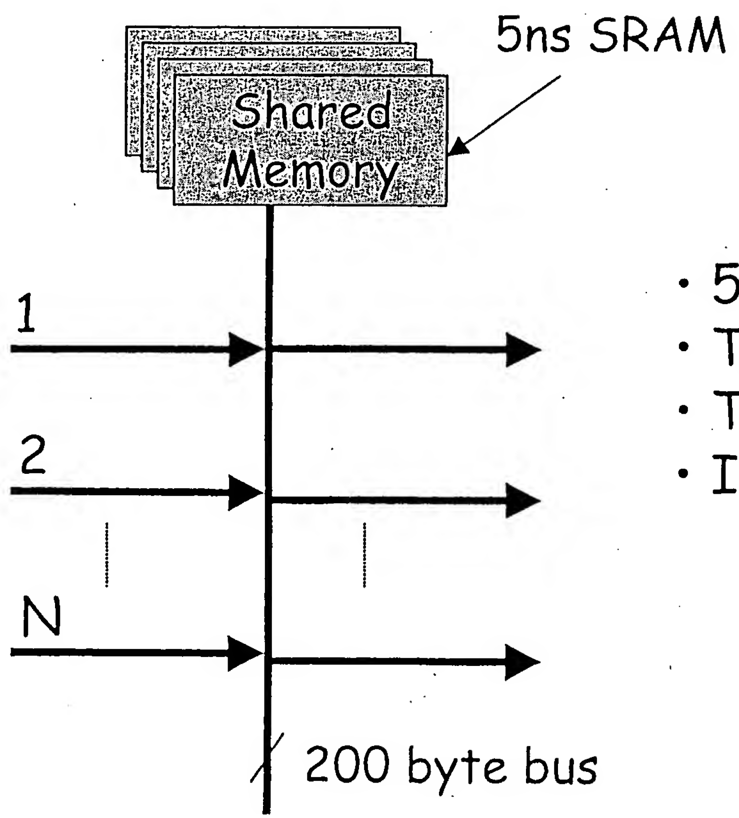
Numerous work has proven and made possible:

- Fairness
- Delay Guarantees
- Delay Variation Control
- Loss Guarantees
- Statistical Guarantees

Large, single dynamically allocated memory buffer:
N writes per "cell" time
N reads per "cell" time.
Limited by memory bandwidth.

Output Queueing

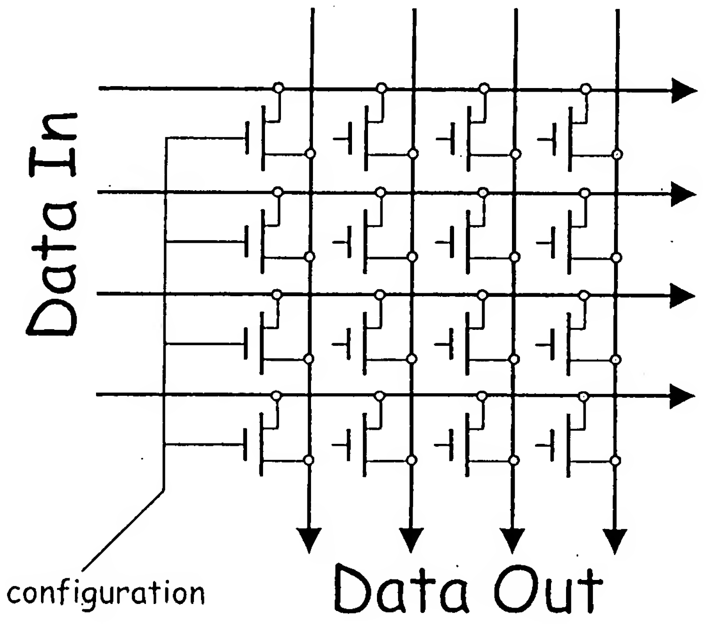
How fast can we make centralized shared memory?

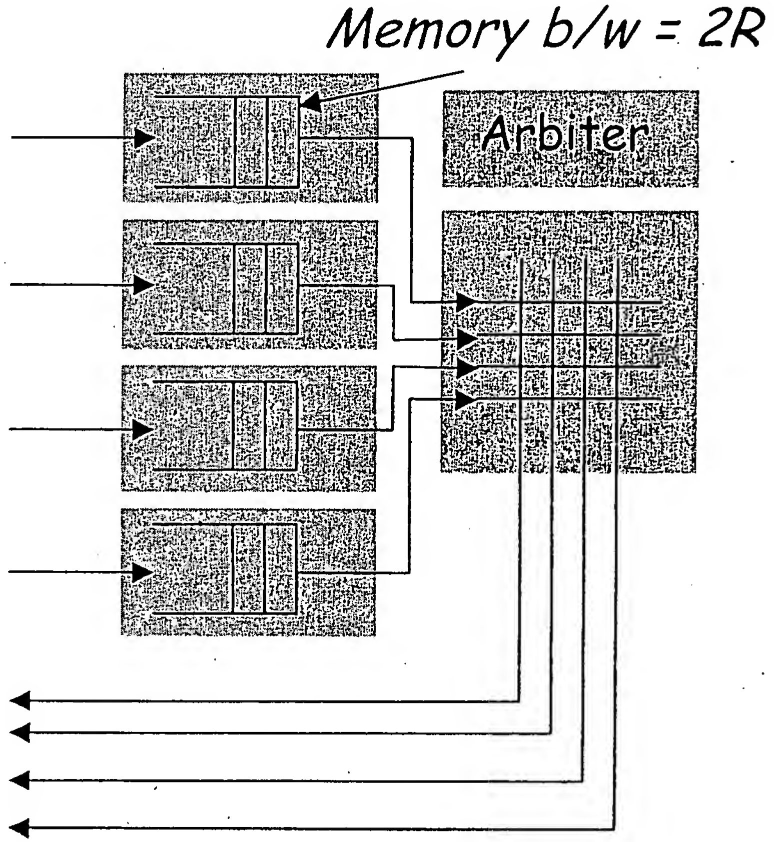


- 5ns per memory operation
- Two memory operations per packet
- Therefore, up to 160Gb/s
- In practice, closer to 80Gb/s

Interconnects

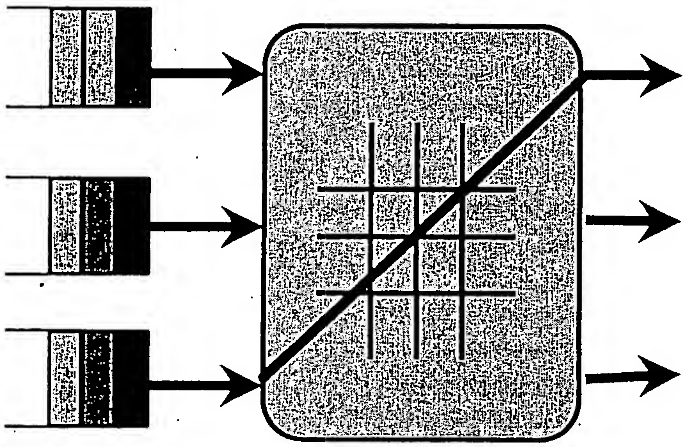
Input Queueing with Crossbar

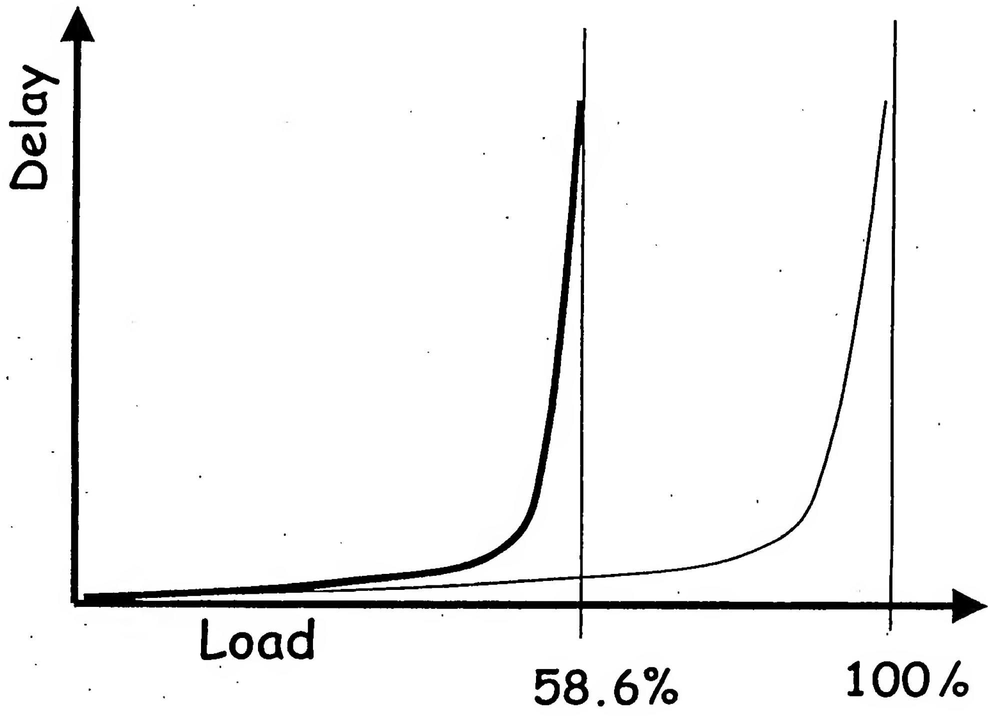




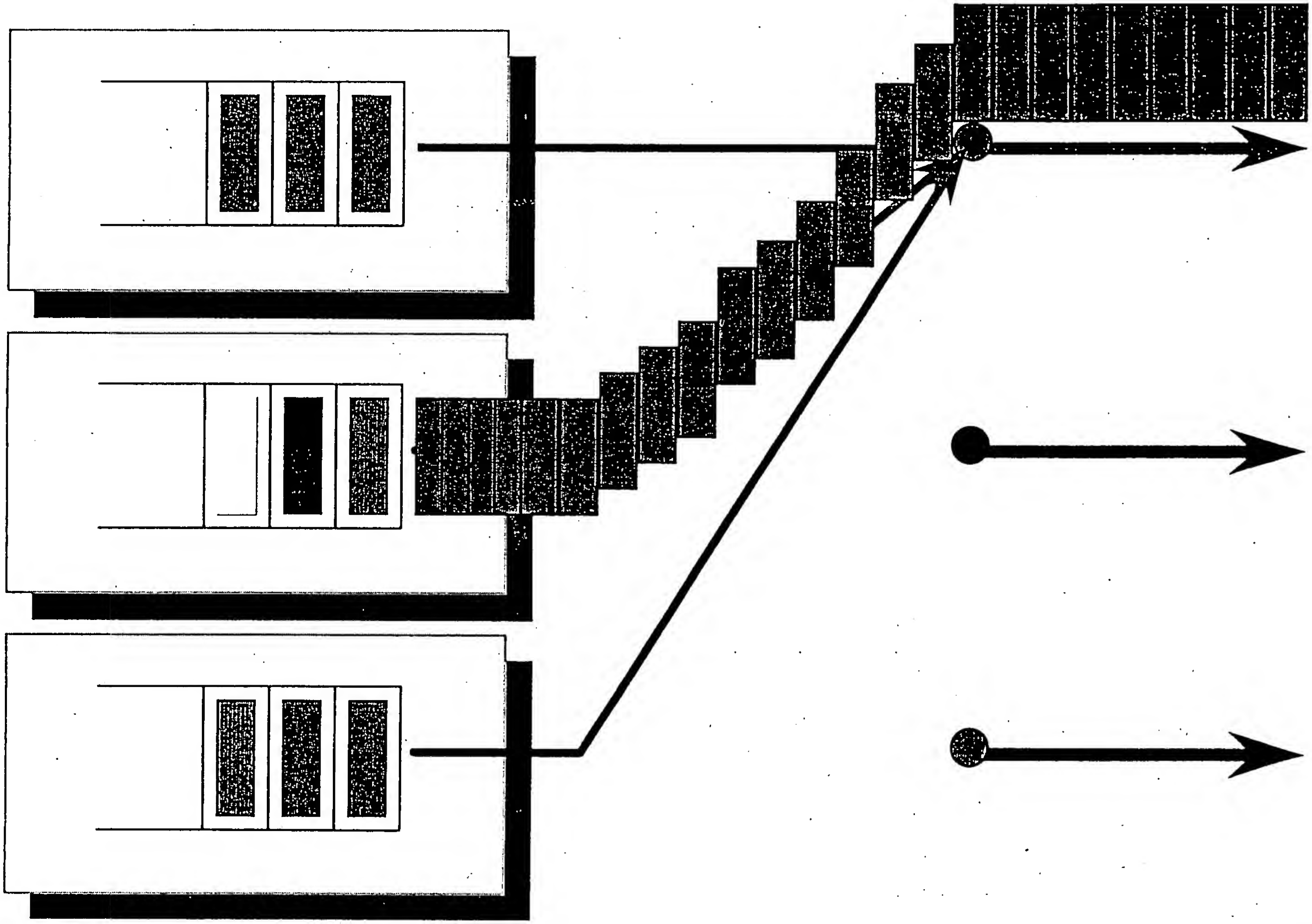
Input Queueing

Head of Line Blocking



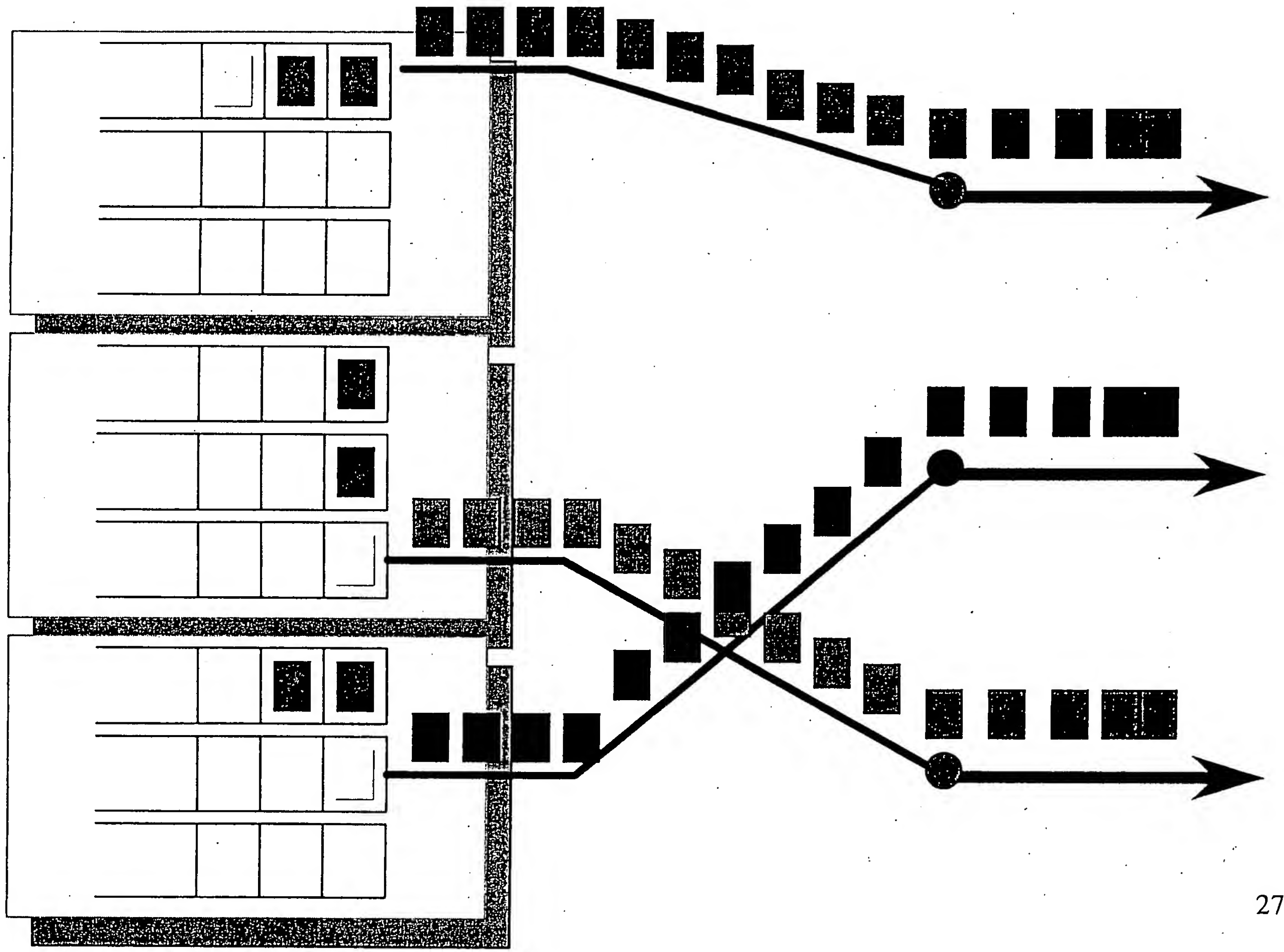


Head of Line Blocking



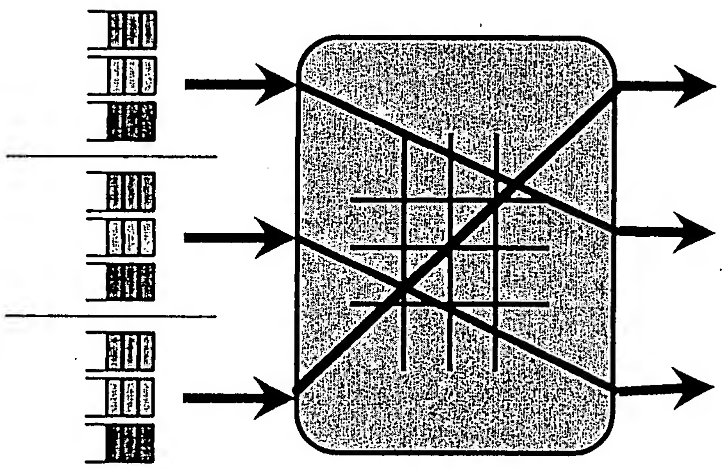
Input Queueing

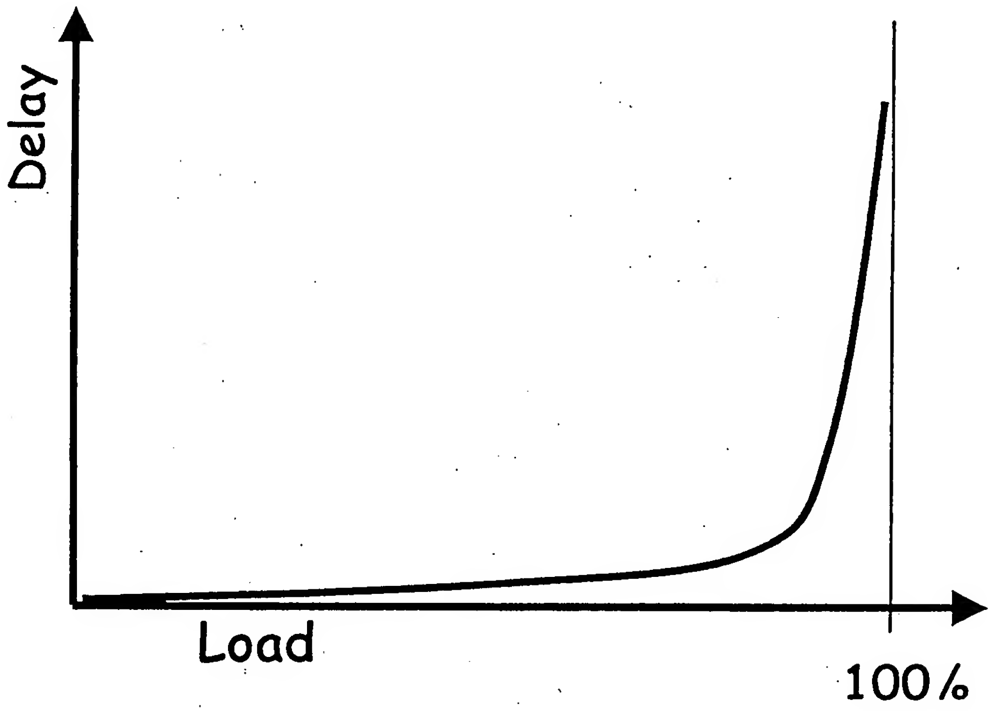
Virtual output queues



Input Queueing

Virtual Output Queues

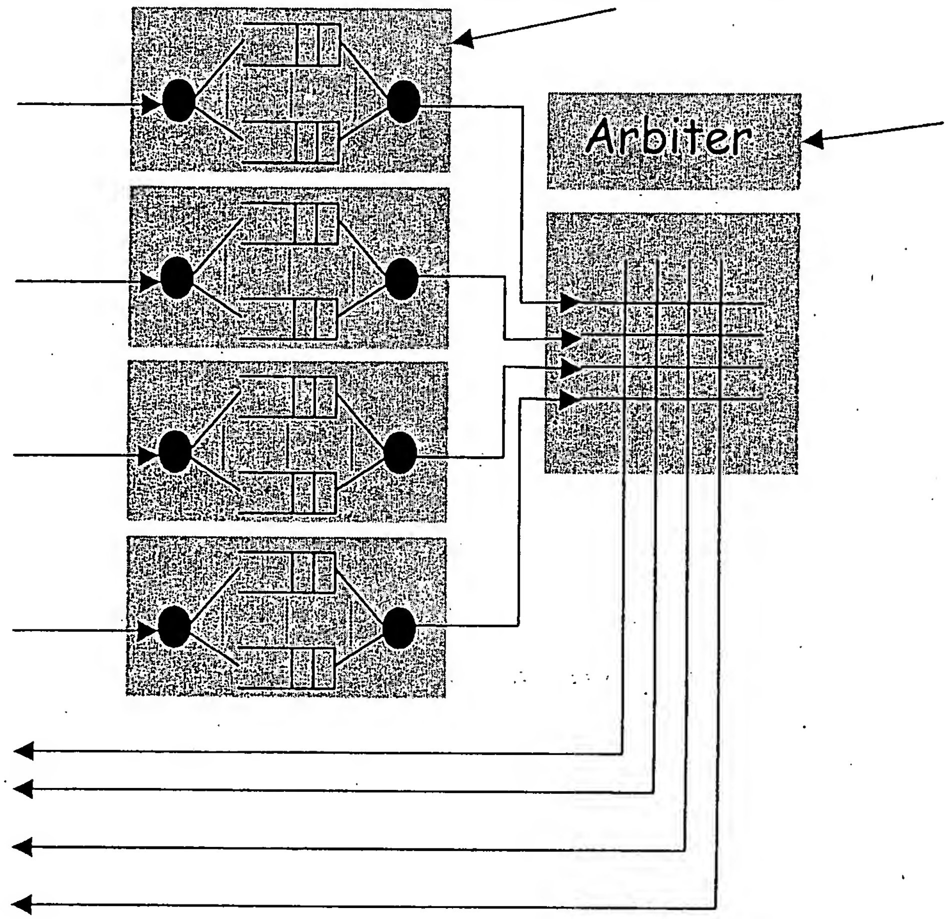




Input Queueing

Virtual output queues

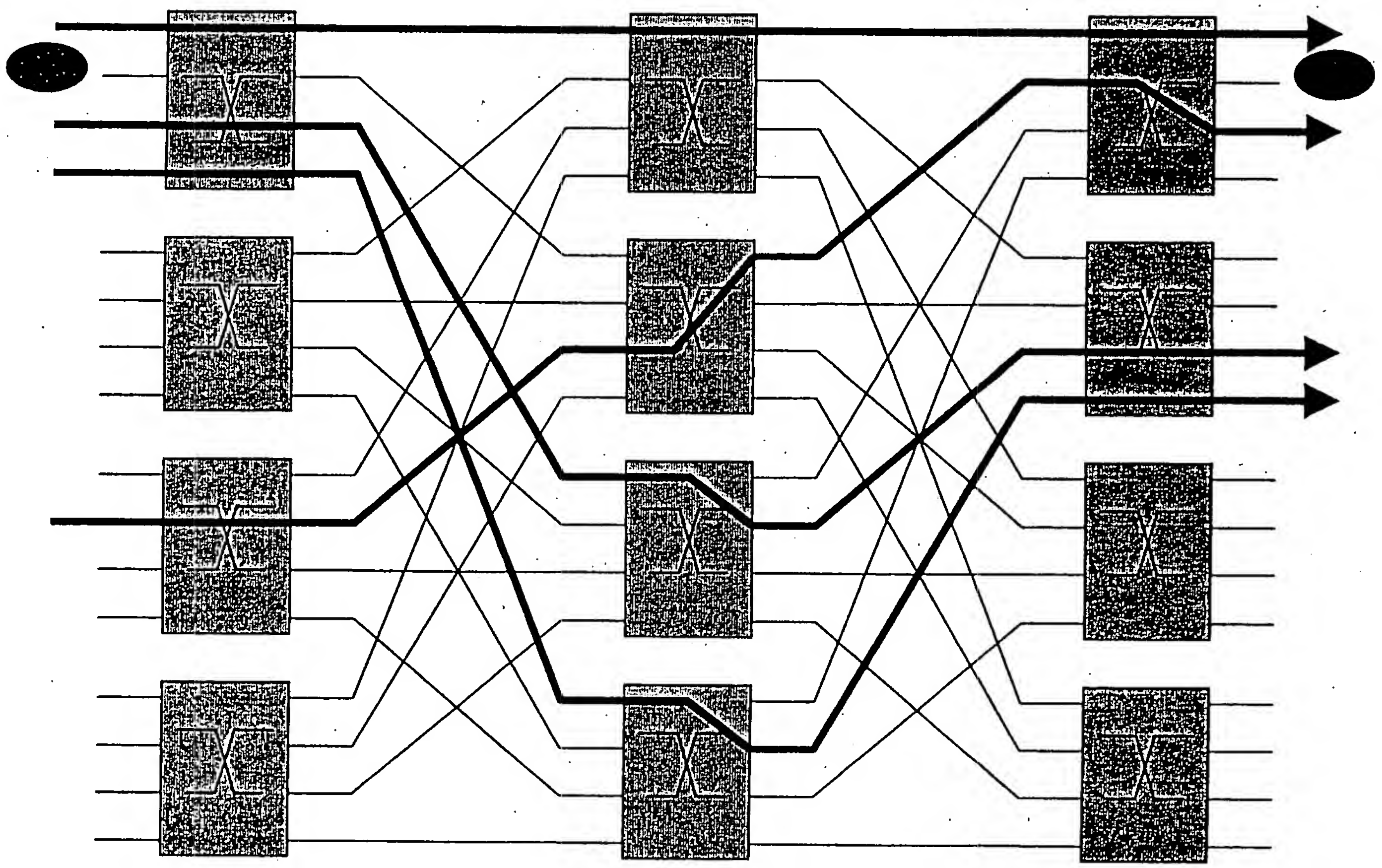
Memory b/w = 2R



Complex!

Other Non-Blocking Fabrics

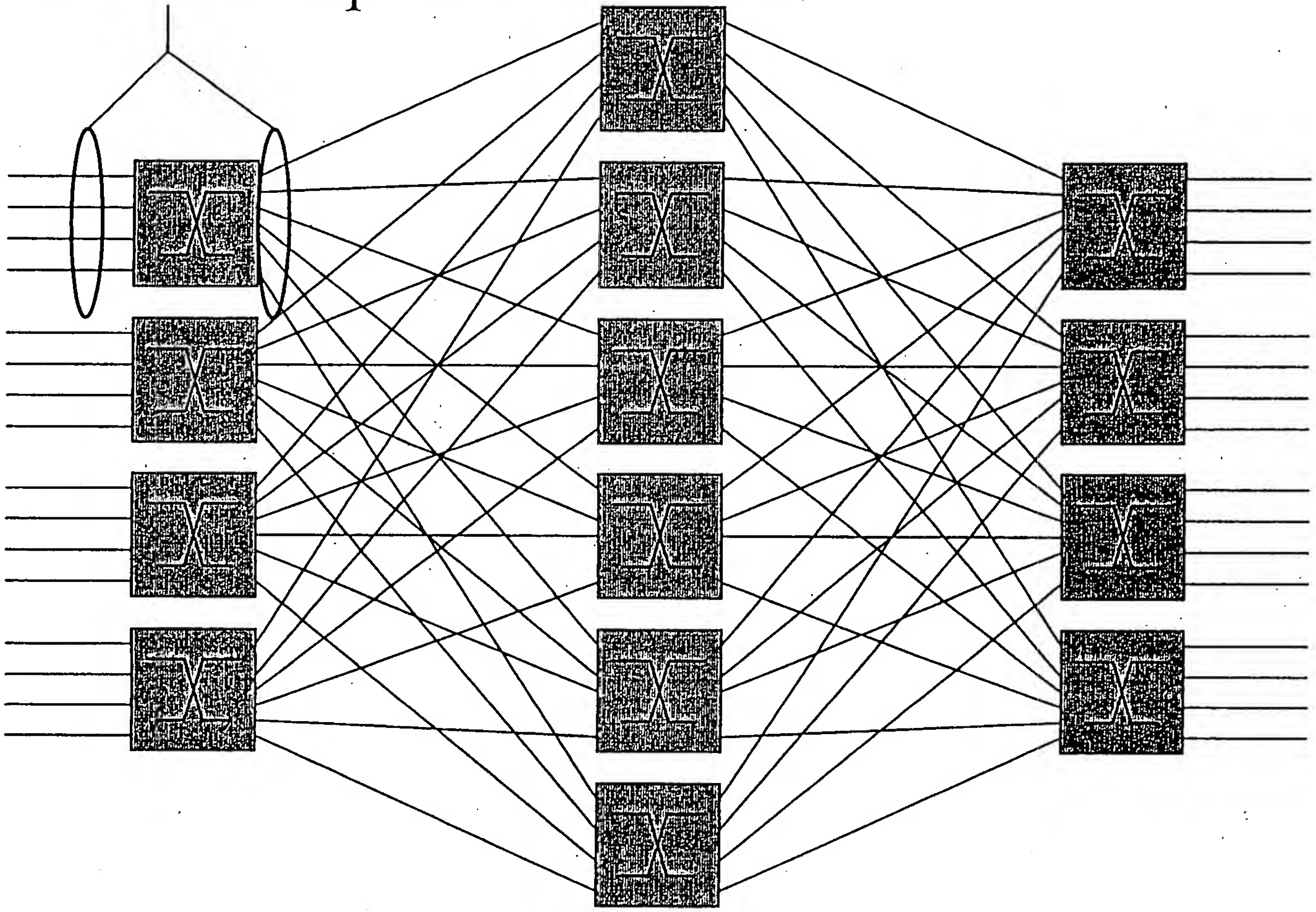
Clos Network



Other Non-Blocking Fabrics

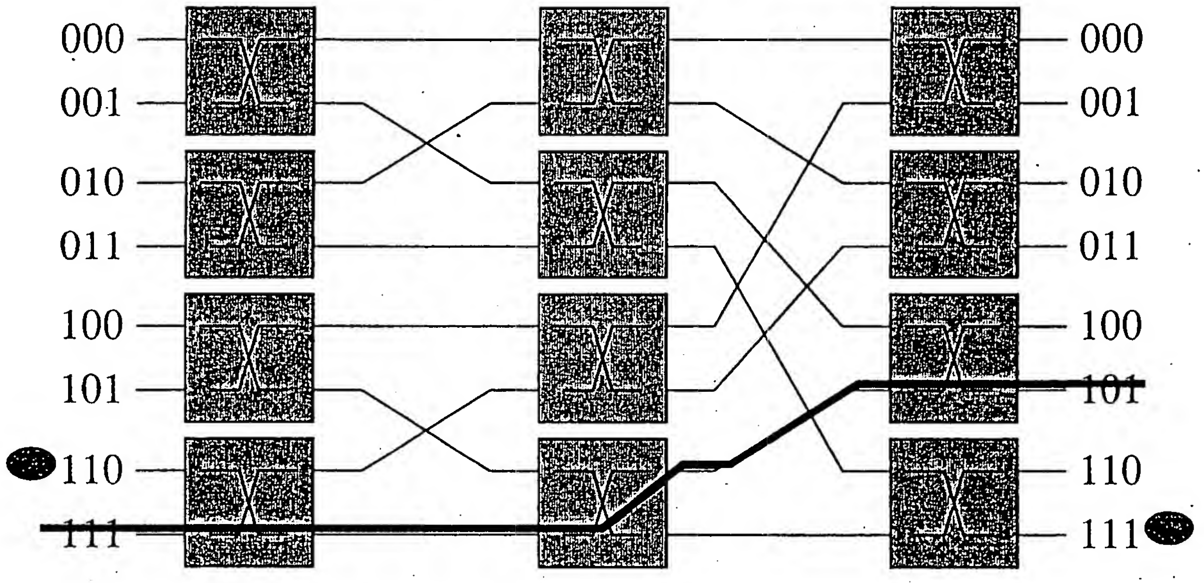
Clos Network

Expansion factor required = 2-1/N



Other Non-Blocking Fabrics

Self-Routing Networks

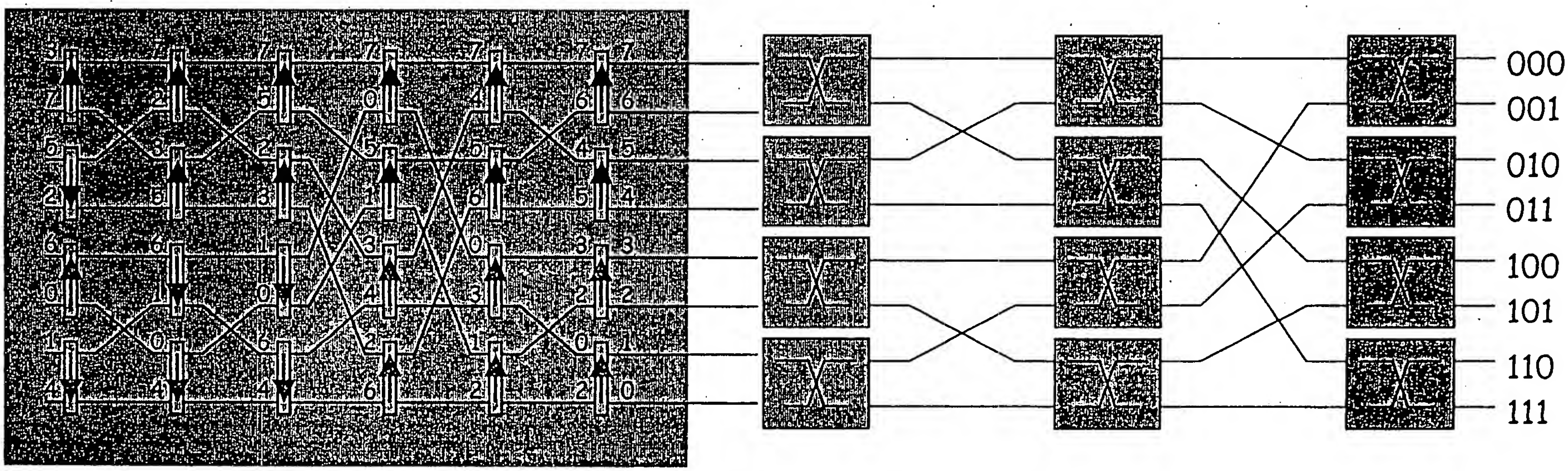


Other Non-Blocking Fabrics

Self-Routing Networks

The Non-blocking Batcher Banyan Network

Bitonic Sorter *Self-Routing Network*



- *Fabric can be used as scheduler.*
- *Batcher-Banyan network is blocking for multicast.*

Outline

Where IP routers sit in the network

- What IP routers look like

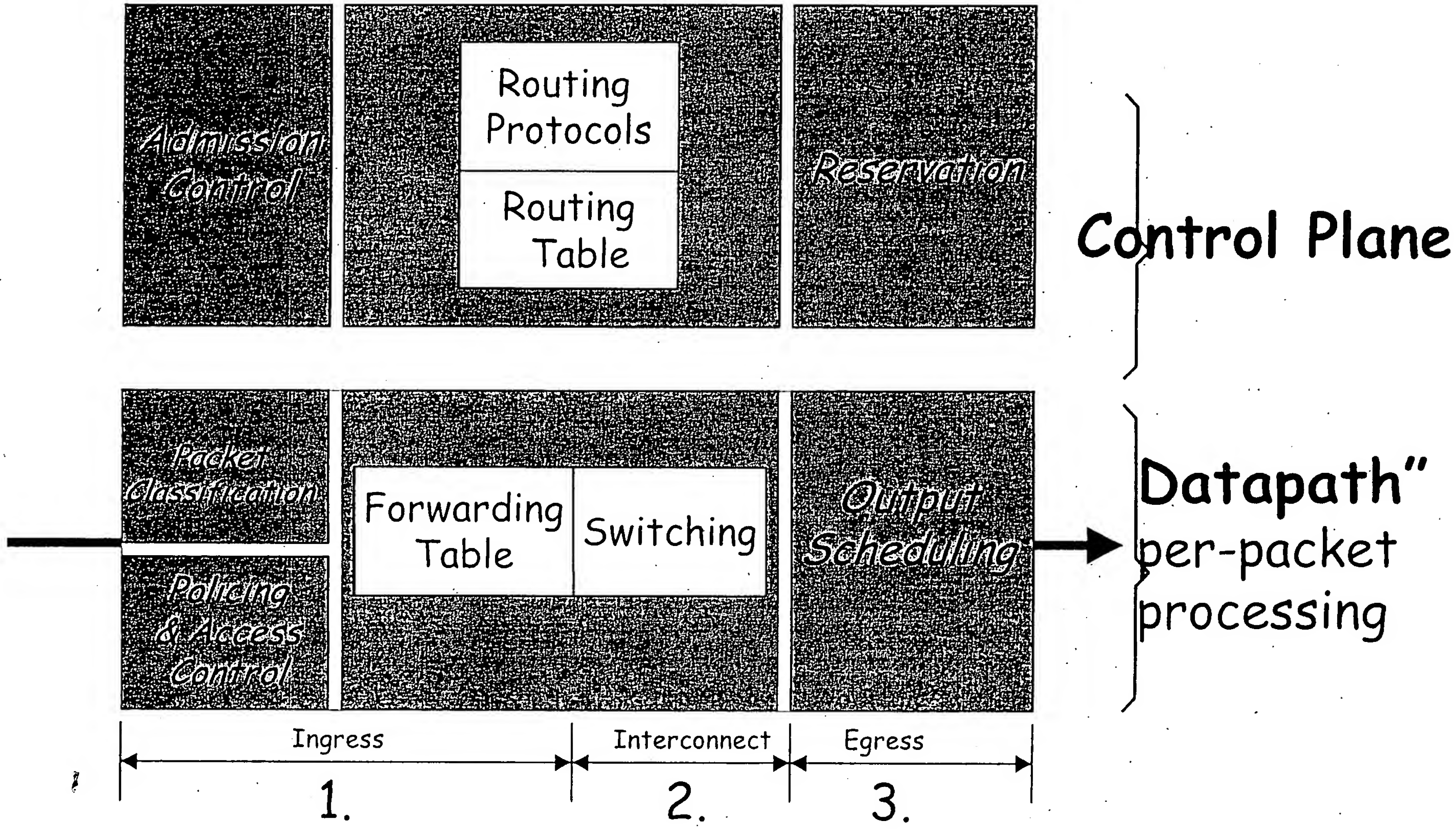
What do IP routers do?

Some details:

- The internals of a "best-effort" router
    - Lookup, buffering and switching
- The internals of a "QoS" router

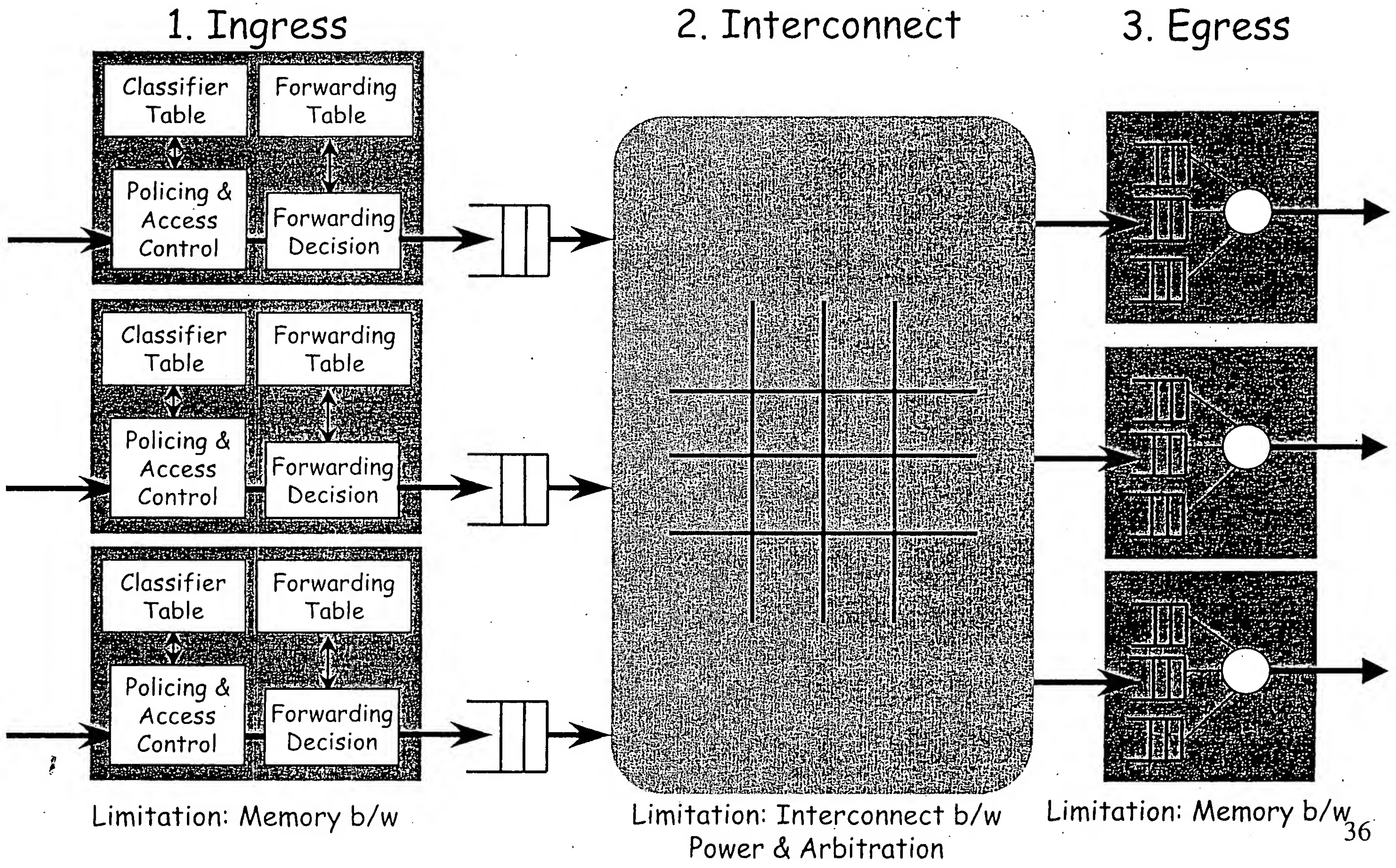
Can optics help?

Basic Architectural Components



Basic Architectural Components

Datapath: per-packet processing



Outline

- Where IP routers sit in the network

What IP routers look like

What do IP routers do?

Some details:

  - The internals of a "best-effort" router
    - Lookup, buffering and switching
  - The internals of a "QoS" router

➡ Can optics help?

Can optics help?

Cynical view:

1. A packet switch (e.g. an IP router) must have buffering.
2. Optical buffering is not feasible.
3. Therefore, optical routers are not feasible.
4. Hence, "optical switches" are circuit switches (e.g. TDM, space or Lambda switches).

Can optics help?

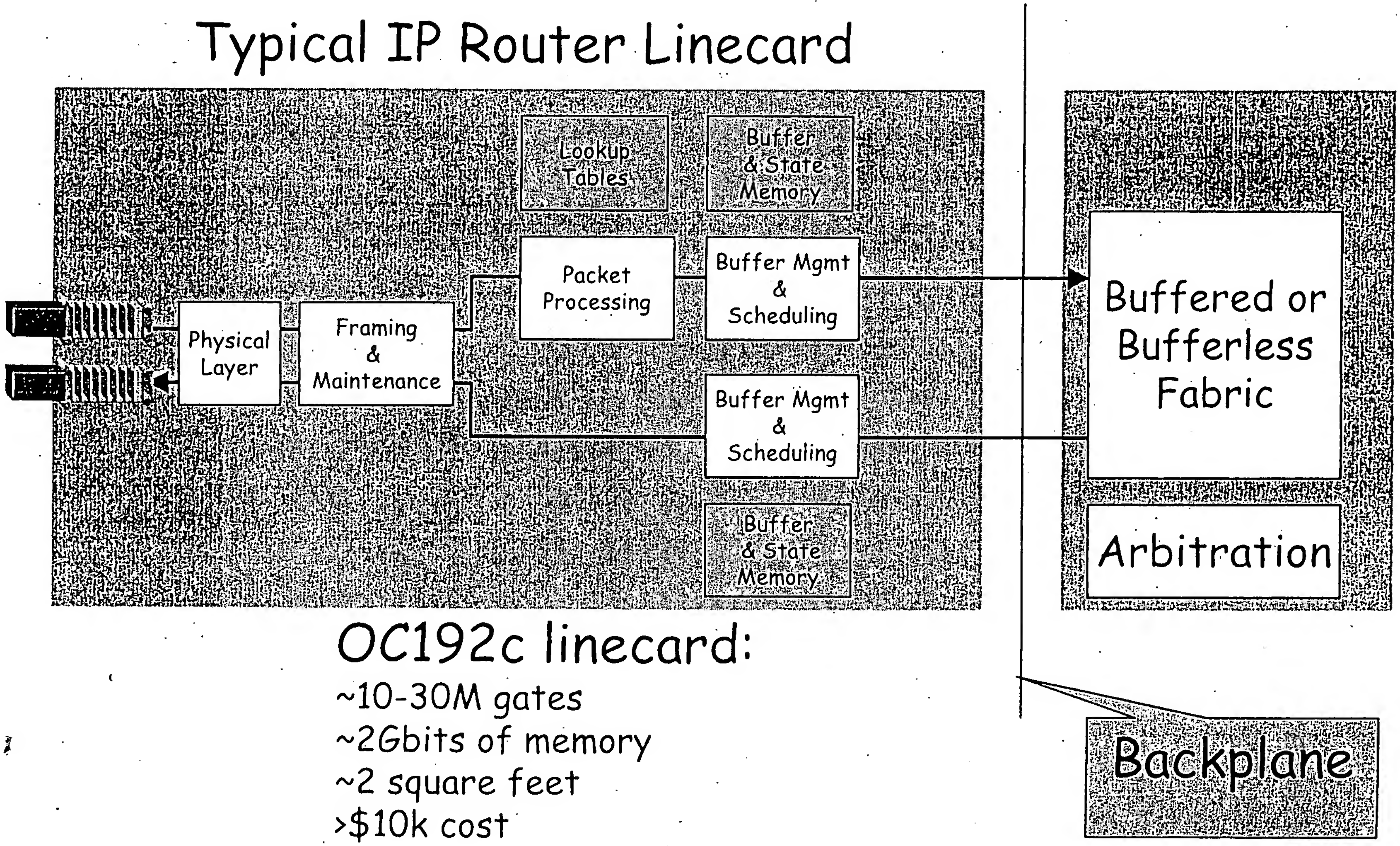
Open-minded view:

Optics seem ill-suited to processing intensive functions, or where random access memory is required.

- Optics seems well-suited to bufferless, reconfigurable datapaths.

Can optics help?

Typical IP Router Linecard



OC192c linecard:

~10-30M gates
~2Gbits of memory
~2 square feet
>$10k cost

Can optics help?

Linecard?

- The linecard is processing & memory intensive.

- Interconnect?
  - Arbitration is very processing intensive.
  - The fabric can be a bufferless datapath...
  - How fast can an optical datapath be reconfigured?

Outline for next time...

The way IP routers are *really* built.

Evolution of their internal workings.

What limits their performance.

The effect that DWDM is having on switch/router design.

The way the network is built today.

Discussion: The scope for optics